

A short introduction to Bayesian statistics, part I

D. Joyce, Clark University

Apr 2009

I'll try to make this introduction to Bayesian statistics clear and short. First we'll look at a specific example, then the general setting, then Bayesian statistics for the Bernoulli process, for the Poisson process, and for normal distributions.

1 A simple example

Suppose we have two identical urns—urn A with 5 red balls and 10 green balls, and urn B with 10 red balls and 5 green balls. We'll select randomly one of the two urns, then sample with replacement that urn to help determine whether we chose A or B .

Before sampling we have “prior” probabilities of $\frac{1}{2}$, that is, $P(A) = \frac{1}{2}$ and $P(B) = \frac{1}{2}$.

Let's suppose our sample $\mathbf{X} = (X_1, X_2, \dots, X_n)$ is of size n , and k of the outcomes X_1, X_2, \dots, X_n are red. We want to determine the “posterior” probabilities $P(A|\mathbf{X})$ and $P(B|\mathbf{X})$, and we'll do that using Bayes' formula. We can easily compute the reverse probabilities

$$\begin{aligned}P(\mathbf{X}|A) &= \left(\frac{1}{3}\right)^k \left(\frac{2}{3}\right)^{n-k} \\P(\mathbf{X}|B) &= \left(\frac{1}{3}\right)^{n-k} \left(\frac{2}{3}\right)^k\end{aligned}$$

so by Bayes' formula we derive the posterior probabilities

$$\begin{aligned}P(A|\mathbf{X}) &= \frac{P(\mathbf{X}|A)P(A)}{P(\mathbf{X}|A)P(A) + P(\mathbf{X}|B)P(B)} \\&= \frac{\left(\frac{1}{3}\right)^k \left(\frac{2}{3}\right)^{n-k} \frac{1}{2}}{\left(\frac{1}{3}\right)^k \left(\frac{2}{3}\right)^{n-k} \frac{1}{2} + \left(\frac{1}{3}\right)^{n-k} \left(\frac{2}{3}\right)^k \frac{1}{2}} \\&= \frac{2^{n-k}}{2^{n-k} + 2^k}\end{aligned}$$

$$\begin{aligned}P(B|\mathbf{X}) &= 1 - P(A|\mathbf{X}) \\&= \frac{2^k}{2^{n-k} + 2^k}\end{aligned}$$

For example, if in $n = 10$ trials we get $k = 4$ red balls, then the posterior probabilities are $P(A|\mathbf{X}) = \frac{4}{5}$ and $P(B|\mathbf{X}) = \frac{1}{5}$.

2 The basic principle

The setting for Bayesian statistics is a family of distributions parametrized by one or more parameters along with a prior distribution for those parameters. In the example above we had a Bernoulli process parametrized by one parameter p the probability of success. In the example the prior distribution for p was discrete and had only two values, $\frac{1}{3}$ and $\frac{2}{3}$ each with probability $\frac{1}{2}$.

Typically, a sample \mathbf{X} is taken from the distribution, and a posterior distribution for the parameters is computed.

Let's clarify the situation and introduce terminology and notation in the general case where X is a discrete random variable, and there is only one discrete parameter θ . In statistics, we don't know what the value of θ is; our job is to make inferences about θ . The way to find out about θ is to perform many trials and see what happens, that is, to select a random sample from the distribution, $\mathbf{X} = (X_1, X_2, \dots, X_n)$, where each random variable X_i has the given distribution. The actual outcomes that are observed I'll denote $\mathbf{x} = (x_1, x_2, \dots, x_n)$.

Now, different values of θ lead to different probabilities of the outcome \mathbf{x} , that is, $P(\mathbf{X}=\mathbf{x} | \theta)$ varies

with θ . In the so-called “classical” statistics, this probability is called the *likelihood* of θ given the outcome \mathbf{x} , denoted $L(\theta | \mathbf{x})$. The reason the word likelihood is used is the suggestion that the real value of θ is likely to be one with a higher probability $P(\mathbf{X}=\mathbf{x} | \theta)$. But this likelihood $L(\theta | \mathbf{x})$ is *not* a probability about θ . (Note that “classical” statistics is much younger than Bayesian statistics and probably should have some other name.)

What Bayesian statistics does is replace this concept of likelihood by a real probability. In order to do that, we’ll treat the parameter θ as a random variable rather than an unknown constant. Since it’s a random variable, I’ll use an uppercase Θ . This random variable Θ itself has a probability distribution, which I’ll denote $f_{\Theta}(\theta) = P(\Theta=\theta)$. This f_{Θ} is called the *prior distribution* on Θ .

The symbol $P(\mathbf{X}=\mathbf{x} | \theta)$ really is a conditional probability now, and it should properly be written $P(\mathbf{X}=\mathbf{x} | \Theta=\theta)$, but I’ll abbreviate it simply as $P(\mathbf{x} | \theta)$ and leave out the references to the random variables when the context is clear. Using Bayes’ law we can invert this conditional probability. In full, it says

$$P(\Theta=\theta | \mathbf{X}=\mathbf{x}) = \frac{P(\mathbf{X}=\mathbf{x} | \Theta=\theta)P(\Theta=\theta)}{P(\mathbf{X}=\mathbf{x})}$$

but we can abbreviate that as

$$P(\theta | \mathbf{x}) = \frac{P(\mathbf{x} | \theta)P(\theta)}{P(\mathbf{x})}.$$

This conditional probability $P(\theta | \mathbf{x})$ is called the *posterior distribution* on Θ . Note that the denominator $P(\mathbf{x})$ is a constant, so the last equation says that the posterior distribution $P(\theta | \mathbf{x})$ is proportional to $P(\mathbf{x} | \theta)P(\theta)$. I’ll write proportions with the traditional symbol \propto so that the last statement can be written as

$$P(\theta | \mathbf{x}) \propto P(\mathbf{x} | \theta)P(\theta).$$

When we discuss the three settings—Bernoulli, Poisson, and normal—the random variable X will be either discrete or continuous, but our parameters will all be continuous, not discrete (unlike the

simple example above where our parameter p was discrete and only took the two values $\frac{1}{3}$ and $\frac{2}{3}$). That means we’ll be working with probability densities instead of probabilities. In the continuous case there are analogous statements. In particular, analogous to the last statement, we have

$$f(\theta | \mathbf{x}) \propto f(\mathbf{x} | \theta)f(\theta)$$

where $f(\theta)$ is the prior density function on the parameter Θ , $f(\theta | \mathbf{x})$ is the posterior density function on Θ , and $f(\mathbf{x} | \theta)$ is a conditional probability or a conditional density depending on whether X is a continuous or discrete random variable.

3 The Bernoulli process.

A single trial X for a Bernoulli process, called a Bernoulli trial, ends with one of two outcomes—success where $X = 1$ and failure where $X = 0$. Success occurs with probability p while failure occurs with probability $q = 1 - p$.

The term Bernoulli process is just another name for a random sample from a Bernoulli population. Thus, it consists of repeated independent Bernoulli trials $\mathbf{X} = (X_1, X_2, \dots, X_n)$ with the same parameter p .

The problem for statistics is determining the value of this parameter p . All we know is that it lies between 0 and 1. We also expect the ratio k/n of the number of successes k to the number trials n to approach p as n approaches ∞ , but that’s a theoretical result that doesn’t say much about what p is when n is small.

Let’s see what the Bayesian approach says here. We start with a prior density function $f(p)$ on p , and take a random sample $\mathbf{x} = (x_1, x_2, \dots, x_n)$. Then the posterior density function is proportional to a conditional probability times the prior density function

$$f(p | \mathbf{x}) \propto P(\mathbf{X}=\mathbf{x} | p) f(p).$$

Suppose, now, that there are k successes occur among the n trials \mathbf{x} . With our convention that

$X_i = 1$ means the trial X_i ended in success, that means that $k = x_1 + x_2 + \dots + x_n$. Then

$$P(\mathbf{X}=\mathbf{x} | p) = p^k(1 - p)^{n-k}.$$

Therefore,

$$f(p | \mathbf{x}) \propto p^k(1 - p)^{n-k} f(p).$$

Thus, we have a formula for determining the posterior density function $f(p | \mathbf{x})$ from the prior density function $f(p)$. (In order to know a density function, it's enough to know what it's proportional to, because we also know the integral of a density function is 1.)

But what should the prior distribution be? That depends on your state of knowledge. You may already have some knowledge about what p might be. But if you don't, maybe the best thing to do is assume that all values of p are equally probable. Let's do that and see what happens.

So, assume now that the prior density function $f(p)$ is uniform on the interval $[0, 1]$. So $f(p) = 1$ on the interval, 0 off it. Then we can determine the posterior density function. On the interval $[0, 1]$,

$$\begin{aligned} f(p | \mathbf{x}) &\propto p^k(1 - p)^{n-k} f(p) \\ &= p^k(1 - p)^{n-k} \end{aligned}$$

That's enough to tell us this is the beta distribution $\text{BETA}(k + 1, n + 1 - k)$ because the probability density function for a beta distribution $\text{BETA}(\alpha, \beta)$ is

$$f(x) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1}(1 - x)^{\beta-1}$$

for $0 \leq x \leq 1$, where $B(\alpha, \beta)$ is a constant, namely, the beta function B evaluated at the arguments α and β .

Note that the prior distribution $f(p)$ we chose was uniform on $[0, 1]$, and that's actually the beta distribution $\text{BETA}(1, 1)$.